Correspondence

Careless talk costs grants

Alan Akers

The editors of *Current Biology* really should be more careful; some contributors are in danger of giving the game away. Years of careful fostering of the idea that everything can be explained in terms of molecular genetics and structural biology — thus keeping many of us gainfully employed — may have been undermined by their careless revelations. The current ascendancy of these disciplines could be at risk if information so casually made available should fall into the wrong hands.

Usually, you play your role exceptionally well. Unqualified assertions that genetic studies of aggressiveness in mice could "…be directly applicable to our understanding of human nature" slipped into the literature without a ripple [1]. Sydney Brenner's suggestion that, because some viral gene products combine in a fixed proportion, "it is possible to encode a mathematical rule in DNA" [2] — with the logical extrapolation that if he, a mere bundle of gene products, jumped out of the window, his DNA would implicitly encode the law of gravity — gave genome studies the opportunity to annex half the funds for chemical and physical research. Then there was the statement about the yeast genome, worthy of a Star Trek script, "…it encourages us to pursue the goal that has been implicit from the beginning: the complete understanding of how a eukaryotic cell functions. The attainment of this lofty goal now seems possible" [3].

Virtual biology and surfing the genomes was all set to abolish the need for messy, wet experiments, which don't always work and, when they do, have a regrettable tendency to throw up untidy, unexpected results. Many of us were looking forward to seeing out our entire research careers with a few clicks on the mouse and a sheaf of publications liberally seeded with comments such as 'intriguing conserved sequence motif', 'could suggest', 'might imply' and (my favourite) 'putative receptor'.

But, just when half the world was convinced that every problem from constipation to criminality is rooted in the base sequence of DNA, you allow loose talk such as "…almost any protein domain can bind inositol phosphate if required" [4] to slip through. Furthermore, the authors openly admit to "…the variety of domains that can bind inositol phosphates", and make things even worse by conceding that "…the functions of most of the binding sites are not yet clear".

Every research scientist using low molecular weight, biologically active compounds soon becomes aware that almost any protein domain can bind almost anything under the right circumstances, but most are instinctively discreet about it. If the promiscuous tendencies of proteins and the sheer variability of biological systems became widely appreciated, ignorant, unscrupulous journalists could call into question our ability to predict the behaviour of living systems from their DNA sequences and protein structures. The fig-leaf of *a priori* reasoning could be shamelessly stripped away to reveal vulgar *post hoc* rationalization.

We can only hope that the relaxed style of *Current Biology* and the speculative license it allows to its contributors have not attracted critical readers from outside the profession.

References

1. Tecott LH, Barondes SH: Genes and aggressiveness. *Curr Biol* 1996, 6:238–240.
2. Brenner S: Molecular biology by numbers …five. *Curr Biol* 1996, 6:490.
3. Johnston M: Genome sequencing: the complete code for a eukaryotic cell. *Curr Biol* 1996, 6:500–503.
4. Irvine R, Cullen P: Inositol phosphates — whither bound? *Curr Biol* 1996, 6:537–540.

Address: Adenauerplatz 8, 69115 Heidelberg, Germany.

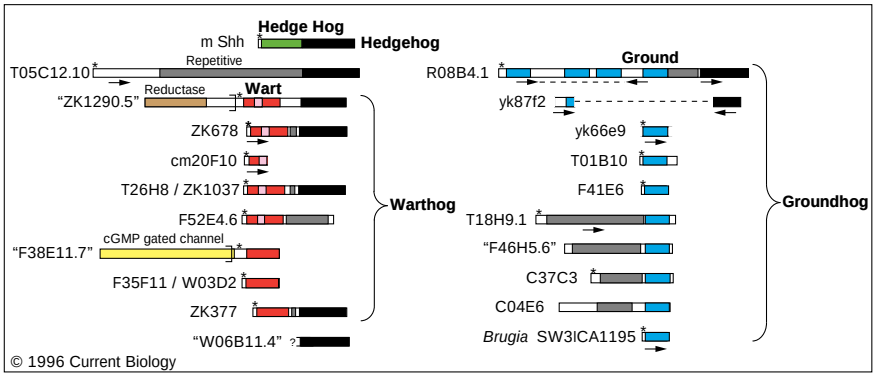
Warthog and Groundhog, novel families related to Hedgehog

Thomas R. Bürglin

Cell–cell signalling is one of the fundamental mechanisms by which different cell fates are generated during development. One group of signalling molecules, encoded by the *Drosophila* gene *hedgehog* and its vertebrate orthologues, has been shown to play important roles during development of flies and vertebrates (see [1–3]). Searching through the *Caenorhabditis elegans* genome, a major fraction of which has now been sequenced [4], reveals several sequences with similarities to *hedgehog* genes. The similarity is restricted to the carboxyl terminus of the Hedgehog proteins, which is surprising given that the amino-terminal part, which provides the biologically active signal, is more highly conserved between fly and vertebrate Hedgehogs. The carboxyl terminus is a distinct domain that has autoproteolytic activity and cleaves Hedgehog into a protease domain and a signalling part [5–7], and it is thought to regulate the release of the amino-terminal signal (see [8]).

The carboxy-terminal domain, which I refer to here as the 'Hog' domain, is about 200–250 amino acids long. Figures 1 and 2a show an alignment of the Hog domains of various Hedgehogs from flies and vertebrates, as well as the predicted products of several of the new *C. elegans* genes. The level of sequence similarity between the *C. elegans* and Hedgehog sequences is of the order of 22–32 % identity in this domain, with highly significant scores produced by BLAST database searches [9]. The probability of a chance match of the ZK678 Hog domain to *Drosophila* Hedgehog is 4.2×10^{-7}. A search of Genbank with the Hog domain has not revealed

Figure 1



The structure of Hedgehog proteins, the *C. elegans* and *B. malayi* Hog, Warthog and Groundhog proteins. The open reading frames (ORFs) used in Fig. 2 were either taken as annotated by the Genome Project or analyzed using Genefinder in ACEDB locally [17]. Local ORF analysis, taking into account the observed sequence similarities, occasionally yielded results different from the annotations by the Genome Project (such ORFs are in quotation marks). In two cases, other genes are inappropriately joined to Warthog genes (ZK1290.5, an aldo/keto reductase, and F38E4.6, a cGMP cation-gated channel protein). In some cases (F46H5 and C04E6) the precise 5′ end could not be determined conclusively. For W06B11.4 the start codon is predicted to be inside the Hog domain; the highly conserved Cys–Phe of the protein splice junction in Fig. 2a has been added manually by extending the ORF towards the 5′ end, and no obvious ORF continuation was found further upstream; perhaps W06B11.4 is a pseudogene. It should be noted that unfinished sequences might harbour mistakes, possibly altering the extent of ORFs outside of the regions of similarity. Hog domains are black, Wart domains red, and Ground domains blue. Protein regions enriched in particular amino acids (labelled 'repetitive', marked with grey boxes) are indicated; they are rich in amino acids such a Gly, Ala, Gln, Pro, Thr and combinations thereof. The arrows underneath the ORFs indicate partial sequences from randomly sequenced cDNAs; some genes are only represented by cDNAs. Asterisks mark signal sequences, and the square brackets indicate where other open reading frames were inappropriately joined by Genefinder.

any other genes encoding this domain.

The autoproteolytic cleavage site in Hedgehog is at a conserved Cys–Phe pair (see arrow, Fig. 2a) within the Hog domain [5], and these two residues are conserved in the *C. elegans* Hog domains. The cleavage site was also predicted by computer comparison to the amino-terminal splice sites of 'intein' protein domains, which are spliced out of larger precursor proteins [10]; the N′-terminal splice junction motif for intein proteins (block A, Fig. 2a) matches all Hog domains. A second weak similarity to intein sequence block B [11] is also observed in the Hog domain (Fig. 2a), but the rest of the Hog domain has no further similarity with inteins.

Novel domains associated with Hog

As the Hog domain is located at the carboxyl terminus (Fig. 1), database searches and comparisons were performed using the amino-terminal parts of the same sequences; these revealed two new motifs. The first, of about 130–170 amino acids, for which I propose the name 'Wart', is present in eight *C. elegans* sequences (Fig. 2b). There appear to be two different types of Wart domain that differ in their central region. Most sequences with a Wart domain also have an associated Hog domain, but three have not (Fig. 1). The Wart domains are highly divergent, but their similarity is nevertheless statistically significant. Their high divergence allows the identification of crucial residues: cysteines appear to play an important role (Fig. 2b). All Warthog sequences have good signal sequences [12] (for protein export) immediately upstream of the Wart domain (Figs 1, 2b).

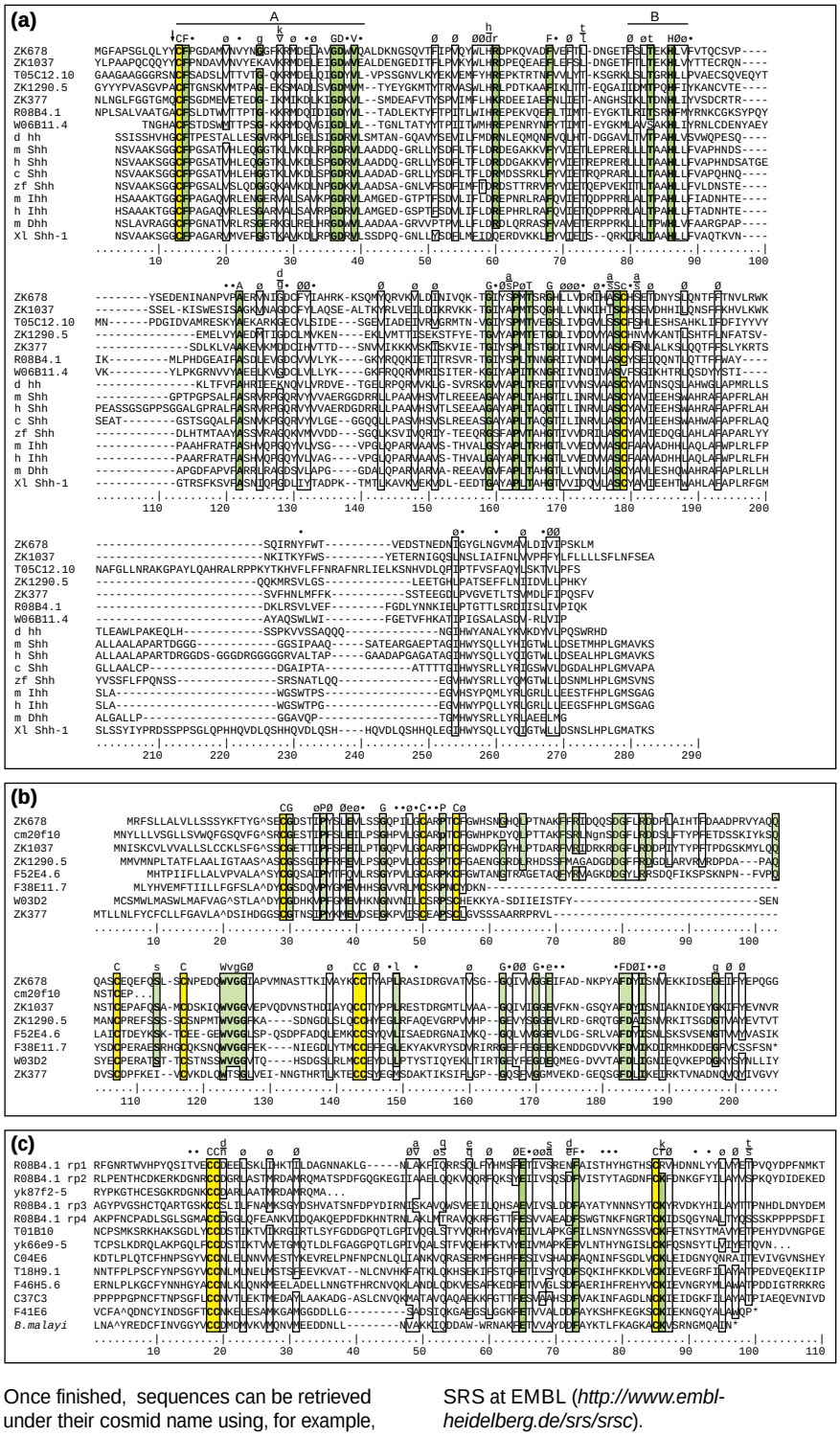
The second motif, for which I propose the name 'Ground', is found in nine *C. elegans* sequences and one from the parasitic nematode *Brugia malayi* (Fig. 2c), estimated to have diverged from *C. elegans* 500 million years ago (D. Fitch and M. Baxter, personal communication). The Ground domain can occur with or without an associated Hog domain (Figs 1, 2c), and some genes contain multiple copies of the Ground domain. As is the case for Wart domains, some Ground domains are highly divergent, most have clearly identifiable signal sequences at their amino termini, and some of the conserved positions are characterized by cysteines. (The amino-terminal 200 amino acids of T05C12.10 show no sequence similarity to anything else in the databases, but it too contains several cysteines and a signal sequence.)

Most of the genes presented here must be functional, as cDNAs have been found for several of them. Because the Hog domain has so far always been found at the carboxyl terminus of a sequence (Fig. 1), it seems unlikely that any protein splicing is involved. The Hog domain might play a role in regulating the release of the amino-terminal signal domain, but several of the Warthogs and Groundhogs do not have a Hog domain, suggesting that it is not always required, and that its putative regulatory role is dispensable under certain conditions.

The fact that most, if not all, of the sequences have a leader motif indicates that they are secreted molecules. In flies and vertebrates, the Hog domain has so far been found only in the Hedgehog proteins, which are involved in cell–cell signalling. The Wart and the Ground domains are both relatively small motifs

characterized by conserved cysteines, reminiscent of other signalling molecules, such as the activin/TGF-β family [13], the NGF-type growth factors [14], the Wnt family [15], and the chordin family [16]. Thus it is possible that the new Hog family members are novel signalling molecules. A member has already been found in a divergent nematode species, so others might occur in flies and vertebrates too. It seems likely

Figure 2



(**a**) The Hog domain. Alignment of the carboxy-terminal regions of various Hedgehog proteins with seven open reading frames (ORFs) from the *C. elegans* genome project. The regions with similarity to intein blocks A and B are indicated at the top. (**b**) Sequence alignment of the Wart domain and associated signal sequences. Lower case letters in cm20f10 indicate uncertain residues which result from corrections of frameshifts and ambiguous nucleotides to maintain protein sequence similarity. The program MACAW [18,19] – available at the EBI worldwide web site (http://www.ebi.ac.uk) – was used to determine statistical probabilities for conserved sequence blocks: the probability of a chance occurrence of the Wart domain is less than 1 in e^{-100}. (**c**) Sequence alignment of the Ground domain. The cDNA yk87f2 shows extended sequence similarity to R08B4.1 outside the Ground domain, so it probably also contains several Ground domains. The probability for the Ground domain as calculated by MACAW, is also less than 1 in e^{-100}. Yellow boxes mark conserved cysteines and green boxes other highly conserved residues. The consensus at particular positions is given on top of the alignments: capital letters indicate an absolute conservation, lower case letters indicate a conservative substitution of that residue. Positions with less conserved residues are marked with open boxes, and preferred residues are given above the alignment. Positions with small hydrophobic residues are indicated by ø for l,L,V,M; small and large hydrophobic residues are marked with a Ø. Positions that are somewhat conserved, displaying only a limited number of residues, are marked by a dot (•). Circumflexes (^) indicate the putative cleavage sites of the signal leader sequences. Accession numbers: *Drosophila melanogaster* (d) hedgehog (hh): L05404; mouse (m) Sonic hedgehog (Shh): X76290; human (h) Shh: L38518; chicken (c) Shh: L28099; zebrafish (zf) Shh: Z35669; m Indian hedgehog (m Ihh): X76291; h Ihh: L38517; m Desert hedgehog (Dhh): X76292; *Xenopus laevis* (Xl) Shh-1: L39213; *B. malayi* SW3ICA1195: N44358. *C. elegans* genes: T05C12: Z66500; ZK1290: U21308; R08B4: Z68008; W06B11: U39854; cm20f10: M89293; F52E4: U56964; F38E11: Z68342; yk87f2: D67840 and D64725; yk66e9: D69245; T19H9: U41746, F46H5: U41543. Unpublished *C. elegans* sequences can be obtained by ftp from *ftp.sanger.ac.uk* (in */pub/C.elegans_sequences/*), and from *http://genome.wustl.edu/gsc/gschmpg.html*. Once finished, sequences can be retrieved under their cosmid name using, for example, SRS at EMBL (*http://www.embl-heidelberg.de/srs/srsc*).

that, in addition to Hedgehog motifs, other motifs associated with a Hog domain will be found in higher animals.

Acknowledgements

I thank J. Groppe, R. Durbin, S. Jones, and L. Hillier for helpful discussions. T.B. is supported by a START fellowship and a grant from the Swiss National Science Foundation.

References

1. Fietz MJ, Concordet J-P, Barbosa R, Johnson R, Krauss S, McMahon AP, Tabin C, Ingham PW: The ***hedgehog*** gene family in ***Drosophila*** and vertebrate development. *Development* 1994, Suppl:43–51.
2. Johnson RL, Tabin C: The long and short of hedgehog signaling. *Cell* 1995, 81:313–316.
3. Perrimon N: Hedgehog and beyond. *Cell* 1995, 80:517–520.
4. Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, *et al.*: 2.2 Mb of contiguous nucleotide sequence from chromosome III of ***C. elegans***. *Nature* 1994, 368:32–38.
5. Porter JA, von Kessler DP, Ekker SC, Young KE, Lee JJ, Moses K, Beachy PA: The product of hedgehog autoproteolytic cleavage active in local and long-range signalling. *Nature* 1995, 374:363–366.
6. Bumcrot DA, Takada R, McMahon AP: Proteolytic processing yields two secreted forms of Sonic hedgehog. *Mol Cell Biol* 1995, 15:2294–2303.
7. Lee JJ, Ekker SC, von Kessler DP, Porter JA, Sun BI, Beachy PA: Autoproteolysis in hedgehog protein biogenesis. *Science* 1994, 266:1528–1537.
8. Bumcrot DA, McMahon AP: Sonic signals somites. *Curr Biol* 1995, 5:612–614.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403–410.
10. Koonin EV: A protein splice-junction motif in hedgehog family proteins. *Trends Biochem Sci* 1995, 20:141–142.
11. Pietrokovski S: Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Protein Science* 1994, 3:2340–2350.
12. von Heijne G: A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res* 1986, 14:4683–4690.
13. Hogan BLM, Blessing M, Winnier GE, Suzuki N, Jones CM: Growth factors in development: the role of TGF-β related polypeptide signalling molecules in embryogenesis. *Development* 1994, Suppl:53–60.
14. Bradshaw RA, Murray-Rust J, Ibáñez CF, McDonald NQ, Lapatto R, Blundell TL: Nerve growth factor: structure/function relationships. *Protein Sci* 1994, 3:1901–1913.
15. McMahon AP: The Wnt family of developmental regulators. *Trends Genet* 1992, 8:236–249.
16. Sasai Y, Lu B, Steinbeisser H, Geissert D, Gont LK, De Robertis EM: ***Xenopus*** chordin: a novel dorsalizing factor activated by organizer-specific homeobox genes. *Cell* 1994, 79:779–790.
17. Durbin R, Thierry Mieg J: A ***C. elegans*** database. Code and data available from anonymous FTP servers *lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk* and *ncbi.nlm.nih.gov 1991-*
18. Schuler GD, Altschul SF, Lipman DJ: A workbench for multiple alignment construction and analysis. *Proteins* 1991, 9:180–190.
19. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993, 262:208–214.

Address: Department of Cell Biology, Biocenter, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland. E-mail: burglin@ubaclu.unibas.ch

Gazetteer

British Biotech plc

What is it famous for? Being a high-profile biotechnology company despite being based in the UK. It was the first biotechnology company to be listed on the London stock exchange; it is now one of about 20.

How did it start? It was founded in July 1986 by two scientists from G.D. Searle & Co., a British pharmaceutical company, when Monsanto bought Searle and promptly closed it. Keith McCullagh, one of the founders, is now the CEO; the other, Brian Richards, was Chairman until the end of 1994.

What does it do? British Biotech's focus from the start has been cancer therapeutics, in particular the matrix metalloproteinases (MMPs) that appear to be important in tissue invasion by malignant cells. Its most promising product at present is an orally active MMP inhibitor, originally given the evocative name BB-2516 but now known as marimastat, which is currently in phase III clinical trials for pancreatic cancer.

Are all the company's eggs in the MMP basket? No. Other drugs in clinical trials include lexipafant, a platelet-activating factor inhibitor now in phase III trials for pancreatitis and phase II for other indications, and BB10010, an engineered version of macrophage inflammatory protein-1α that is intended to protect stem cells during chemotherapy (currently in phase II trials). Phase II trials of p24-VLP, a "virus-like-particle" and candidate HIV therapeutic vaccine, proved "less than encouraging" and were dropped.

How is the company funded? Mostly by stock sales; there is no income from drugs as yet. It recently launched a rights issue to raise £143 million (about $212 million), thought to be the largest amount ever raised by a biotech company. Less than half of the shares were taken up by existing shareholders, however; this was disappointing, but not fatal, as underwriters made up the difference.

Why did it need to raise more money? Although British Biotech's cash reserves stood at around £66 million before the rights issue, the company plans extensive (and expensive) clinical trials for both of its lead products in the near future. The company claims that this should be the last time it needs to appeal to the stock market for cash before the money from hoped-for drug sales starts coming in. The company also wants to consolidate its UK operations at a new site in Cowley, Oxford.

Does the company have any presence in the US? Yes. British Biotech Inc. was founded in 1993 in Annapolis. It's responsible for clinical trials in the US, and for dealing with the US Food and Drug Administration.

Why are there so few biotechnology companies in the UK? British Biotech's success in raising money suggests that the problem is not one of funding. The lack of entrepreneurial spirit in the 'old country' is a popular explanation, as always; the fact that venture capitalists are less aggressive outside the US may also be a factor.